Analyzing NOAA Storm Damage Data

Project Proposal

Matthew Tuttle
CSPB
University of Colorado Boulder
CO USA
matu8468@colorado.edu

Lex Bukowski
CSPB
University of Colorado Boulder
CO USA
Alexi.Bukowski@colorado.edu

ABSTRACT

Severe weather events have a profound impact on society and the world. Storms can destroy infrastructure and property, shut down critical systems, and endanger human lives. The National Oceanic and Atmospheric Administration (NOAA) has tracked and documented severe weather events, the damage they cause, and the number of fatalities dating back to 1950. The following project proposal outlines a data mining and analysis exercise that endeavors to answer interesting questions about the NOAA Storm Event data.

The end goal of the project is to enhance our understanding of severe storms and the damage they cause to humans and property by looking at any possible correlations that may exist in the data. The proposed project will utilize data mining techniques to uncover value in the massive amount of data that the NOAA has collected over the past 70+ years. By implementing methods such as clustering, classification, and regression, this project will provide storm forecasters, community leaders and decision makers, and everyday citizens with key information about storm events and the damage and destruction that they may bring.

CCS CONCEPTS

• Mathematics of computing • Mathematics of computing ~ Probability and statistics • Mathematics of computing ~ Mathematical analysis • Mathematics of computing ~ Probability and statistics ~ Probabilistic representations • Mathematics of computing ~ Probability and statistics~ Probabilistic inference problems • Mathematics of computing ~ Probability and statistics ~ Probabilistic algorithms

KEYWORDS

Storm events, Storm damage, Property damage, Natural disaster, Weather events, Severe weather, Weather damage, Weather fatalities

ACM Reference format:

Lex Bukowski and Matthew Tuttle. 2023. Analyzing NOAA Storm Damage Data. *4 pages.*

1 Problem Statement and Motivation

Many interesting questions about the severe weather events and storm damage data exist such as: What type of weather is the most dangerous to people and property? Is the magnitude of property damage indicative of higher risk of injury to people? Are certain locations prone to multiple types of dangerous weather events? Do longer weather events lead to more property damage? Have severe weather events increased over time?

Answering these questions will allow for a better understanding of the risk of severe weather events and allow for better prediction of the outcomes of severe weather events. Armed with this information, future weather events, even if more severe, can cause less damage to people and property, potentially saving lives and preventing huge expenditures.

The questions surrounding the occurrence of events over time will give insight into possible other causation factors as well as other interesting questions to be explored. Most notably if severe weather events are increasing over time, does

global warming and climate change play a role in the occurrence and severity of these events.

2 Literature Survey

In conducting background research on this project, topic, and data set, two previous published papers were discovered. They are discussed below.

2.1 Text Mining of NOAA Data

Emma Louise McDaniel analyzed the NOAA dataset previously by implementing a text mining algorithm to uncover the true nature and severity of a severe weather event. The logic supporting this study was that although important for insurance adjustments, monetary cost is not always the most prevalent or usable metric to determine the scale and severity of a storm event. McDaniel summarizes this argument with the following quote from her abstract "The monetary impact of a disaster is not conducive for disaster preparedness planners to build community resilience."[1]

McDaniel's work was subsequently to text mine the narratives given about the severe weather events included in the database. This work proved to be difficult as the natural language was very unstructured. However, that doesn't make the information contained in the narrative any less important. McDaniel's work was a first step toward extracting and classifying key words and terms out of the storm event narratives to determine a relationship between those key terms and a storm's true (not just monetary) effect. Below is a citation of McDaniel's work, which is also included in the references section at the end of this proposal.

SAC '23: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. March 2023 Pages 653 – 656 https://doi.org/10.1145/3555776.3577211

2.2 Comments on Reliability of NOAA's Storm Events Database

Renato P Dos Santos analyzed the reliability, completeness, and high-level validity of the dataset. Dos Santos understood the limitations of validating every input into the database due to the size of the dataset, but took a high-level approach to do a quick analysis on completeness and consistency of the data.

Dos Santos concluded that although there have been standardization efforts on the part of the NOAA to make the database more usable and effective, there are many non-standard event types in the database. Along with those inconsistencies, Dos Santos added that many damage reports are missing and many damage values are incorrect.

Below is a reference to Renato P Dos Santos work, which is also included in the references section at the end of this proposal.

P dos Santos, Renato, Some Comments on the Reliability of NOAA's Storm Events Database (June 22, 2016). Available at SSRN: https://ssrn.com/abstract=2799273 or http://dx.doi.org/10.2139/ssrn.2799273

3 Data Set

The data set that this project will work with is the "NOAA Storm Events Database." This database is free and open to the public and can be found at the following URL: https://www.ncdc.noaa.gov/stormevents/ftp.jsp.

Summarized but detailed information about the contents of the database can be found at the same URL. The database includes columns that contain information on the details, locations, and fatalities of each storm event, linked by the primary key of the event ID number.

The Details table includes information such as the timing, general location, event type, reported injuries, reported deaths, crop damage, property damage, and the storm narrative (more information is housed in this table, these are just the most relevant fields for this proposal). The Storm Data Location table includes much more detailed location information and the Storm Data

Fatality table does the same for the reported fatalities.

It is worth noting that the database has entries from three different phases. From 1950-1954, the database contains only information on tornado damage. From 1955-1995, damage from thunderstorms, wind, and hail was also included. Finally, from 1996 onward, data was collected for a total of 48 event codes. Care will have to be taken to constrain analysis of events over time to the proper time periods based on the available data.

4 Proposed Work

4.1 Preprocessing

There is a significant amount of preprocessing that will need to be completed on this data set. The heart of the dataset are the fields EVENT_TYPE, PROPERTY_DAMAGE, CROP_DAMAGE, INJURIES_DIRECT, and INJURIES_INDIRECT. All of these fields have some significant issues that will need to be addressed before the data can be analyzed.

The EVENT_TYPE field officially has 48 possible event types, but the database includes a total of 70 different labels. Many of these non-standard labels are combinations of existing labels. It will need to be determined if the non-standard labels need to be recategorized, and how combination labels should be dealt with.

The PROPERTY_DAMAGE and CROP_DAMAGE fields represent storm damage in dollar value. However, these are not numeric fields but instead are stored as strings with a magnitude signifier as the last character. For example, damage of $25,000 might be stored as '25.0K' with K signifying a magnitude of 1,000. There are a number of non-standard signifiers (e.g. '?') which will likely result in the data rows being removed. Additionally, there are a significant number of entries in the data set that have empty values for both damage types. Given that inclusion in the data set is predicated on damage this is likely an error and will need to be resolved.

Finally, the INJURIES_DIRECT and INJURIES_INDIRECT fields appear to be much less problematic as they are integer valued. However, only 0.7% of all data entries caused either a direct or indirect injury. Additionally, over half the total injuries in the database were caused by 5 events. With such sparse/concentrated data, conclusions will have to be drawn very carefully.

4.2 Data Integration

Population and infrastructure density are a significant variable in storm damage that is not currently captured in the database. The database does provide standardized region labels for storm locations as well as the latitude and longitude of storms. If needed, these fields can be used to tie in an outside data source to provide per capita normalization of damage and injuries.

4.3 Classification and Analysis

After completing the data preprocessing and any required or added data integration, classification and analysis will be performed on the cleaned and complete data set to determine possible correlations that exist in the dataset.

4.4 Conclusion

The goal of this project is to make informed conclusions about correlations between severe weather events occurrence, damage, and death toll by analyzing the other factors provided in the data set. The last step of the project will be drawing these conclusions from the data processing, integration, and classification steps. Care will be taken to not apply unnecessary causation, but simply show determined correlations.

A possible conclusion for this project, as noted from previous works done above, is that the data set, although encompassing over 70 years of information, does not contain enough structured information to determine interesting correlations with a strong support basis.

5 Evaluation Methods

The possible correlations between attributes of the data set and/or if the occurrence of attributes is

increasing (or decreasing) with time will be evaluated using various methods. Graphically, information from scatter plots and box plots will be used along with more numerical methods of R-squared and Pearson correlation coefficients.

Due to previously referenced work completed by P dos Santos, evaluation of the data set as a whole needs to be addressed as well. The data set will be analyzed for completeness and consistency while evaluating correlations. Confidence intervals and residual analysis will be utilized to help determine if correlations appear poor due to poor data quality or accurate data analysis.

6 Tools

A number of tools will be used for this analysis. An exploratory analysis of the data will be performed with python using pandas and numpy paired with matlibplot and seaborn for data visualization.

Classification and regression analyses will be done in python using a combination of the sklearn and statsmodel libraries.

Documentation for all described libraries is readily available online.

6.1 Exploratory Tools

Pandas and Numpy are libraries built on top of the Python language that will help simplify, clean, and analyze the dataset. Both tools will be used extensively with the goal of standardizing and gaining a high level understanding of the data. This will allow for effective classification and regression analysis, as opposed to choosing data traits randomly. Matlibplot and seaborn will help visualize the information gained during this data scrubbing process and will also provide visuals for final presentation of the data.

6.2 Classification and Regression Tools

The sklearn and statsmodel libraries are tools built on top of python that will allow for effective classification and regression on data traits that were identified as important and relevant from the cleaning phase.

7 Milestones

Data preprocessing will be completed the week of November 13th.

Data integration will be completed the week of November 27th.

Data classification and analysis will be completed the week of December 4th.

Conclusions, final reports, and presentations will be completed the week of December 11th.

ACKNOWLEDGMENTS

Acknowledgements should be made to both Emma Louise McDaniel and Renato P Dos Santos whose previous work with the data was taken into consideration when determining the proposed work and evaluation methods included in this proposal.

Professor Kristy Peterson (University of Colorado Department of Computer Science) and the University of Colorado CSPB 4502 Fall 2023 course participants also provided feedback, critiques, and suggestions to this proposal idea and draft.

REFERENCES

[1]SAC '23: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. March 2023 Pages 653 – 656 https://doi.org/10.1145/3555776.3577211

[2]P dos Santos, Renato, Some Comments on the Reliability of NOAA's Storm Events Database (June 22, 2016). Available at SSRN: https://ssrn.com/abstract=2799273 or http://dx.doi.org/10.2139/ssrn.2799273